barcelona|gse
graduate school of economics

14D009 - Social and Economic Networks

Project Report

Random walk based centrality measures

Author: Felix Gutmann
Student number: 125604
Program: M.S. Data Science
E-Mail: felix.gutmann@barcelonagse.eu

I Table of Contents

II List of Figures

III List of Tables

IV List of mathematical symbols

Symbol	Meaning
δ	Kronecker delta

1 Introduction and review of relevant papers

Various measures have been proposed to find important nodes in a network. However, the term importance depends on the question one might to answer. The most natural way to think about importance is how many connection a node has, called the *degree centrality* of a node. Besides that two very prominent measures were proposed to take the global structure of the network into account.
Closeness Centrality indicates how close a node is on average to all other nodes in the network. Therefore, it measures how quickly a node interacts with the rest of the network. [Wasserman and Faust, 1994, page 184].
On the other hand *betweness centrality* counts how often a node lies on the shortes path [Wasserman and Faust, 1994, page 190]. Hence, it can be interpreted as the power a node has on the flow of information in a network [Newman, 2005, page 2]. Both of the latter concepts assume that information is passed along the shortest path between two nodes and hence are suitable to answer question where information a transmition happens in such a way. However, that assumption might not always be valid. necessarily be the case for obvious reasons. People might discuss different topics with different kind of friends and thus information spread has a different nature. Hence one might also take no geodesics into account. Another point is that connection might change or disappear. Errasing a node from a network (e.g a friendship ends) can change the outcome for the mentioned centrality measure. Therefore, random walk based measures provide a more robust centrality indicator [Mavroforakis et al., 2016, page 1]
One way to adress this issue is to use random walks on graph. This first project provides a summary of relevant work in that field and give a formal derivation of proposed concepts. Therefore, the report has the following structure. First there will be an informal review of the papers outlining the intention of the concepts and possible fields of applications. The following section then provides a formal derivation of each concept. In the following we will review the following four paper. To my knowledge [Noh and Rieger, 2004] and [Newman, 2005] can be viewed as the two defining papers in that field. [Mavroforakis et al., 2016] is a recent paper

doing also something. While the later three defining random walks in the notion of *markow chains*, [Takes and Kosters, 2011] approaches random walks in a different way.

2 Mathematical foundation of centrality measures

This section provides formal derivations for each centrality measure. Since, most papers use different notations we might define some basic notation to ensure consistency. In general we try to merge the notation of the papers following [Mavroforakis et al., 2016].

2.1 Notation and basic backround of markow chains

First we might formally introduce a graph. Note, that we only consider unweighted and undirected simple graphps. A graph is an ordered pair $G = (V, E)$, where V is the set of vertices $V = \{v_1, \dots\}$ and E is the set of edges $E = \{e_1, \dots, e_n\}$. The adjacency matrix $\mathbf{A}$ of the graph G [Aigner, 2007, page 124], where an entry is defined as:

$$a_{ij} = \begin{cases} 1 & \text{if } \mathrm{v}_i\mathrm{v}_j \in \mathrm{E} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The *neigbherhoud* of a node $N(v_i)$ is the set of vertices adjacent to v_i. Hence, the *degree* of a node is defined as:

$$d(v_i) = |N(v_i)| \tag{2}$$

Using equation (2) we define $\mathbf{d}$ as the n $\times$ 1 vector of all degrees and $\mathbf{D}$ as the diagonal matrix of the degrees.

In general a random walk is characeterized as follows. A random walk starts at a given node of the graph and moves to an adjacent node with a certain probability. They are usually modeled with the notion of markow chains. We introduce alson some notations for the markow chains following [Wasserman, 2004, page 383 et.

seqq.]. Consider the set of different states $\mathcal{X} = \{1, \dots, N\}$. A markow chain is modeled with an N $\times$ N *transition matrix* $\mathbf{P}$, where an entry describes the probabilty to go move from one state to another. A random walks is a stochastic process and therefore developes over a a number of (in our case) discrete steps denoted by $T = \{1, 2, \dots\}$. Combining that, an entry of the transition matrix satisfies the following:

$$\mathbf{P} = p_{ij} := \mathbb{P}\left(X_{t+1} = j | X_t = i\right) \tag{3}$$

The next step is therefore only defined by the last state. The transition probabilities in the for step t is obtained by simple matrix multiplication:

$$\mathbf{P}(t) := \underbrace{\mathbf{P} \times \cdots \times \mathbf{P}}_{\text{t - times}} \tag{4}$$

A state is called *persistent* if $p_{ii} = 1$ and $p_{iij} = 0$ for $j \neq$ and *transient* otherwise. Hence persisent can be seen as absorbing nodes. This fact will be applied later in section 2.4.

2.2 Random walk centrality

The paper especially introduces the concept of the *Mean first passage time*.[1] We start by defining the transition probabilities for the random walk by normalizing the the entries of the adjacency matrix by the degree.

$$\mathbf{P} = \mathbf{A}\mathbf{D}^{-1} \tag{5}$$

Note that we deviate a little bit from the source paper and continou to express the necessary derivation in matrix algebra. We might normalize each degree by the sum over all degress. Let $D = \sum_k \mathbf{d}_i$.

$$\mathbf{P}_i^{\infty} = \frac{d_i}{D} \tag{6}$$

$$p_{ij}(t) = \delta_{t0}\delta_{ij} + \sum_{t'=0}^{t} p_{jj}(t - t')F_{ij}(t') \tag{7}$$

[1] The concept gets for example adapted and adjustet to compute central sectors in Input Output Tables (for more information see [Blöchl et al., 2010] and [Blöchl et al., 2011])

2.3 Random walk betweness centrality

The tranisition matrix is identical to equation (9). We also want to see the transition behaviour. Hence we could set the values in the transition matrix, such that the target note becomes persistent (this is applied in next section). However, another method is just to remove the t row and column. Hence we denote this matrix as:

$$\mathbf{P}_t = \mathbf{A}_t \mathbf{D}_t^{-1} \tag{8}$$

2.4 Absorbing random walk centrality

For this measure we may have to introduce some additional notation. The objective is to to indentify a set of k central node. Denote this set of central nodes as C. Moreover, define the *query nodes* Q such that $Q \subseteq V$. We define a third set of so called *candidate nodes*, which are potentially in C, such that $C \subseteq D$. This distinction serves the purpose that depending on the application the set of central nodes can be limited to the nodes in Q, but in others can be potentially belong to V.

The random walk in this concept starts at a node from the set in the query nodes $q \in Q$. The random walks proceeds as soon it arrives at a node $c \in C$, where it gets absorbed. The starting node is chosen with a discrete distribution $s(v_i)$.[2]

Likewise the other centrality the centrality is defined as the expected value how fast that happens. Recall the tranistion matrix defined in equation (3). Since a node in C absorbs the random walk it is persistent and the probiltiy of escaping is zero (see the properties of markow chains in section 2). The remaining transient nodes are therefore T = V \ C.

An extension to the algorithm is that a random walk can be restarted with a given probability α. Taking the last mentioned facts we write out the adjusted transition probabilities for

$$p_{ij} = \begin{cases} \alpha s(v_j) & \text{if } j \in Q \setminus N(i) \\ \frac{(1-\alpha)}{d_i} + \alpha s(v_j) & \text{if } j \in N(i) \end{cases} \tag{9}$$

Finally the complete transition matrix can be expressed as a blockwise arrangement of the following four sub matrices:

[2] In the simplest case this can be taken as a uniform distribution

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_{\mathrm{TT}} & \mathbf{P}_{\mathrm{TC}} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \tag{10}$$

Recalling ones more equation (4) we the probability that the random work after t has't been absorbed is $\mathbf{P}_{\mathrm{TT}}(t)$. The expected total number can be computed by the infinite series:

$$\mathbf{F} = \sum_{t=0}^{\infty} \mathbf{P}_{\mathrm{TT}}(t) = (\mathbf{I} - \mathbf{P}_{\mathrm{TT}})^{-1} \tag{11}$$

Using the previous equation (13) that the expected length of the random walk getting absorbed can be computed using the following vector:

$$\mathbf{L} = \mathbf{L}_C = \begin{pmatrix} \mathbf{F} \\ \\ \mathbf{0} \end{pmatrix} \mathbf{1} \tag{12}$$

The final measure is achieved by summing over all nodes in Q.

$$C_{RWA} = \mathbf{s}^{\mathrm{T}} \mathbf{L}_C \tag{13}$$

However, this measure comes with a cost. Not only follows from equation (13) that each computation has to be done inverting a matrix, which can be computationally expensive, but also we have to optimize this procedure over different choices of C.

2.5 Guided random walk centrality

As stated in section 1. this paper has a slightly different approach. First we might define what they call the neighbherhood density:

$$f_{nd}(v) = 1 - \frac{1}{|N(v)|} \tag{14}$$

$$f_{nd}(v) = 1 - \sum_{w \in N(v)} \frac{|N(w) \cap N(v)|}{(|N(w)| - 1)|N(v)|} = 1 - \sum_{w \in N(v)} \frac{|N(w) \cap N(v)|}{(\mathrm{d}_w - 1)\mathrm{d}_v} \tag{15}$$

$$\mathbb{P}(w) = \frac{\alpha f_{deq}(w) + (1-\alpha) f_{nd}(w)}{\sum_{u \in N(v)} (\alpha f_{deq}(u) + (1-\alpha) f_{nd}(u))} \tag{16}$$

3 List of Literature

[Aigner, 2007] Aigner, M. (2007). *Discrete Mathematics*. American Mathematical Society, Providence.

[Blöchl et al., 2010] Blöchl, F., Fisher, E. O., and Theis, F. (2010). Which Sectors of a Modern Economy are most Central? *CESifo Working Paper Series*, (3175).

[Blöchl et al., 2011] Blöchl, F., Theis, F. J., Vega-Redondo, F., and Fisher, E. O. (2011). Vertex Centralities in Input-Output Networks, Reveal the Structure of Modern Economies. *Phys. Rev. E*, 83(4).

[Mavroforakis et al., 2016] Mavroforakis, C., Mathioudakis, M., and Gionis, A. (2016). Absorbing random-walk centrality: Theory and algorithms. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2016-Janua:901–906.

[Newman, 2005] Newman, M. E. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54.

[Noh and Rieger, 2004] Noh, J. D. and Rieger, H. (2004). Random Walks on Complex Networks. *Physical Review Letters*, 92(11):118701–1.

[Takes and Kosters, 2011] Takes, F. W. and Kosters, W. A. (2011). Identifying prominent actors in online social networks using biased randomwalks. *Belgian/Netherlands Artificial Intelligence Conference*.

[Wasserman, 2004] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*, volume 1542.

[Wasserman and Faust, 1994] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, Cambridge.

Appendix